

A stochastic model of the tweet diffusion on the Twitter network

Tatsuro Kawamoto

*Department of Physics, The University of Tokyo, Komaba, Meguro, Tokyo 153-8505,
Japan*

Abstract

We introduce a stochastic model which describes diffusions of *tweets* on the Twitter network. By dividing the followers into generations, we describe the dynamics of the tweet diffusion as a random multiplicative process. We confirm our model by directly observing the statistics of the multiplicative factors in the Twitter data.

Keywords: Twitter, social network, random multiplicative process, information diffusion

1. Introduction

As a popular microblogging web service, a significant attention is paid to Twitter. One of the important points for users of Twitter is how well one's writings, or *tweets*, are spreading through the network. Its significance is obvious from the facts that the counters which measure the popularity of tweets are everywhere on the web and that many companies are using Twitter as an advertising tool. There have already been many data related to Twitter [1, 2, 3, 4, 5, 6, 7] presented in the literature.

On Twitter, users can *follow* other users and read their tweets without any approvals, thereby constructing a directed network among them. Users can also propagate tweets to their followers thanks to a characteristic function called *retweet* [8], which results in information diffusion. The aim of the present paper is to introduce a stochastic model for the tweet diffusion of the *daily tweets* on the Twitter network. The reason why we investigate the daily tweets is because we can expect a universal behavior; in the case where we selected tweets with specific keywords, we might observe irregular behaviors

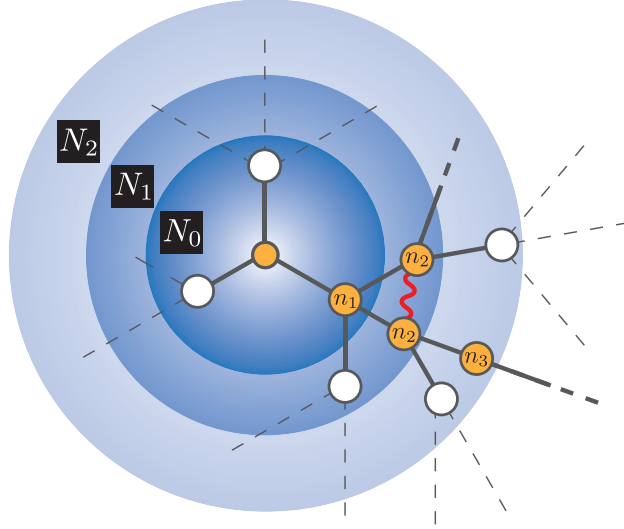


Figure 1: (Color online) Diffusion network on Twitter. The node at the center represents the seed account and the linked nodes are the followers. A solid line means that the tweet has diffused through the link by the retweet. In the model analysis, we ignore the over-counting of followers such as the one illustrated by the wavy line, while we take account of it in the data analysis.

depending on the characteristics of the keywords. As we will see, any tweet diffusion of the daily tweets seems to have a random multiplicative process as the underlying mechanism.

We believe that the information diffusion on Twitter is different from some other web contents which are discussed in the literature. It has been known that the total spread of the information of many web contents obey lognormal distributions [9, 10, 11]. There, one can think of a discrete-time random multiplicative process; the additional number of spreads in one time step depends on the total number of spreads by the previous time steps and a decaying function of time, *i.e.*, $N_{t+1} = N_t(1 + r_t X_t)$, where t is the discrete time, N_t is the total number of spreads by time t , X_t is a positive stochastic variable, and r_t is the decaying function. It is a natural modeling for featured contents and contents that people search for. In the case of Twitter, however, it is unnatural to assume such a mechanism because we expect that tweets diffuse through the followers and the probability of the retweet activity should not depend on the total spread; users usually do not look for the tweets to retweet. They retweet because they receive the tweets.

In the present paper, we will propose to classify the followers into *generations* depending on the distance from the seed account and consider a stochastic process along them. The diffusion process which we present would probably occur in many other networks, especially on the web. An advantage of researching the Twitter data is that we can confirm our model directly thanks to the Twitter API [12], which provides rich information.

The present paper is organized as follows. In Sec. 2, we will introduce a stochastic model of the tweet diffusion along the generations. In Sec. 3, we will confirm that our model is indeed plausible by directly observing the stochastic variables of the model in the Twitter data. Note that we will fix a seed account when we analyze the data and there are some restrictions for the selection of the seed account (see Sec. 3.1). Finally, after summarizing the present paper, we will argue what can be further expected in Sec. 5.

2. Model

Before we construct a model, let us first explain how a tweet diffuses by retweet on Twitter in detail. Figure 1 shows a schematic picture of the tweet diffusion process. Whenever a user generates a tweet, it will be sent to N_0 followers of the tweet owner, whom we call users in the zeroth generation. Next, when n_1 users out of N_0 followers retweet, the original tweet will be sent to the followers of the n_1 retweeters; we call them users in the first generation. We label the number of the receivers in the first generation as N_1 . Such a chain of diffusion of a tweet continues until people stop retweeting or all the followers in the last generation are the users who already received the tweet [13]. We will refer to the total number of receivers as $N_{\text{tot}} = \sum_{g=0}^{\infty} N_g$ and the total retweet count as $n_{\text{RT}} = \sum_{g=1}^{\infty} n_g$, where g stands for the label of the generation. While N_0 is simply the number of the followers of the seed account, N_g for $g \geq 1$ reads

$$N_g = \sum_{f=1}^{n_g} k_f - c_g, \quad (1)$$

where f stands for the label of each retweeter and k_f stands for the number of his or her followers. The factor c_g is the number of over-counting of the followers (*e.g.*, the wavy line in Fig. 1). In the case where the network is close to the tree structure and the distribution of the number of followers is

homogeneous, we can employ the approximation

$$N_g \simeq n_g \sum_{k=0}^{\infty} k p_g(k) =: n_g \bar{k}_g, \quad (2)$$

where k and $p_g(k)$ are the number of the followers of the retweeters in the $(g-1)$ th generation and its distribution, respectively.

Let us next estimate the number of the retweeters, n_g . Since there are N_{g-1} candidates to generate the retweeters in the g th generation, we assume

$$n_g = \beta_g N_{g-1}, \quad (3)$$

where β_g is a variable which we call the retweet rate. Although β_g is a discrete variable because n_g and N_{g-1} are integers, we treat it as if it were a continuous variable. In Sec. 3, we will observe that the retweet rate has a distribution over many incidents of tweet diffusion. We therefore regard β_g as a continuous stochastic variable hereafter.

Combining eqs. (2) and (3), we have

$$N_m = J_m N_{m-1} = \cdots = \prod_{g=1}^m J_g N_0, \quad (4)$$

$$n_m = \beta_m N_{m-1} = \cdots = \beta_m \prod_{g=1}^{m-1} J_g N_0, \quad (5)$$

where

$$J_g = \beta_g \bar{k}_g, \quad (6)$$

which is a stochastic variable because β_g is a stochastic variable. Although the probability distribution of J_1 may strongly depend on the characteristics of the seed account, J_g for $g \geq 2$ are expected to obey a common probability distribution. Therefore, the number of viewers of the tweet in each generation, N_g , is expressed as a random multiplicative process because of the hierarchical structure of the followers. Note that the present model is not a standard percolation model, which assigns a stochastic variable to each follower, but a stochastic process with respect to each generation. We do not consider the time dependence of the retweet rates since most of the tweets finish diffusing very quickly [2].

In the following section, we will directly observe the statistics of the retweet rates β_g and confirm that our modeling is indeed plausible.

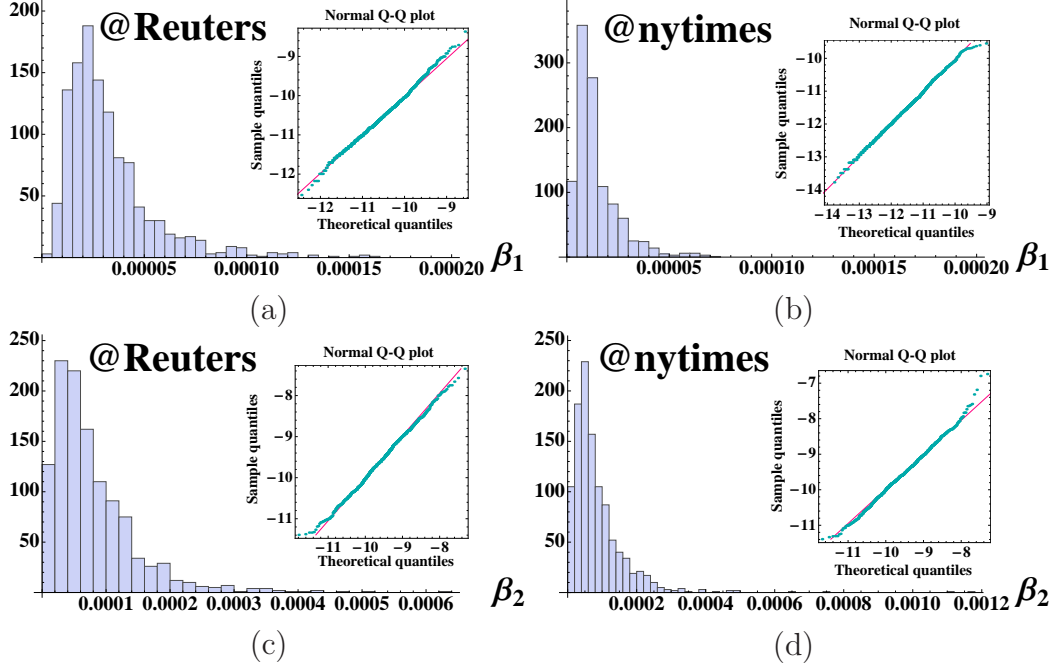


Figure 2: (Color online) The histograms of the retweet rates and the normal Q-Q plots of the logarithms of the retweet rates. (a) and (c) are of β_1 and β_2 for @Reuters. (b) and (d) are of β_1 and β_2 for @nytimes.

3. Data analysis for the retweet rate β_g

Using the data sampled by the tool Twitter API [12], we directly observed the behaviors of β_1 and β_2 . We chose The New York Times (@nytimes) and Reuters Top News (@Reuters) for the seed accounts and sampled the diffusion data with $n_2 > 0$. The data are summarized in Table 1.

3.1. Possible errors, selection of the seed accounts, and restrictions

There are some inevitable errors in our data. We cannot sample the data of private users and there might be some miscounts in n_g because the follow-followed relation might have changed by the time we sampled the data. In order to sample the data as accurately as possible, we need to select the seed accounts carefully; we chose the seed accounts which tweet frequently and the number of whose followers are not changing rapidly so that we can expect the network around the seed account is almost static during the period of sampling. In order to see the statistical behavior clearly, it is good to choose

an account with a large number of followers and high retweet rates [14]. When we analyze the retweet rate β_g , we take into account the factor of over-counting c_g in Eq. (1), and thus we do not assume a tree structure nor the homogeneity of the distribution of the followers.

3.2. Result

Figures 2(a) and 2(b) show the histograms of β_1 and their normal Q-Q plots [15]. They show that the retweet rate β_1 seems to obey lognormal distributions with slight additive shifts, *i.e.*

$$\beta_1 = e^{\omega_1} + \delta_1, \quad (7)$$

where ω_1 obeys Gaussian distributions $\mathcal{N}(\mu_1, \sigma_1^2)$ with μ_1 being the mean and σ_1^2 being the variance of $g = 1$. For β_1 , the mean μ_1 and the variance σ_1^2 seem to depend strongly on the character of the seed account. The slight additive shift might be due to the systematic activities by Twitter bots.

We expect that the retweet rate β_2 also obeys lognormal distributions with slight additive shifts. Figures 2(c) and 2(d) show the histograms of β_2 and their normal Q-Q plots; they indeed indicate the lognormal behaviors. For β_2 , the mean μ_2 and the variance σ_2^2 are very close for both of the seed accounts; it seems to be plausible to model that the retweet rate β_g obeys a common probability distribution for $g \geq 2$.

In Table 1, we listed the averages of the over-counting of the users in the first generation, *i.e.* $c_1 / \sum_{f=1}^{n_1} k_f$ in Eq. (1). The over-counting of users are less than 5% on average, and thus the networks around the seed accounts have almost the tree structures. Although it is still doubtful whether the tree-structure approximation is appropriate in all generations, it is hard to imagine a drastic qualitative change to the diffusion phenomenon due to the loop correction since there is no back flow.

Since we are fixing the seed account, N_0 is a constant and the distribution of β_1 is proportional to that of n_1 . The number of followers in the first generation, N_1 , and the number of retweeters among them, n_2 , can take different values for each sample. As are shown in Figs. 3(c) and 3(d), both of them obey lognormal distributions and they are not independent of each other. The correlation coefficients of n_2 and N_1 , $\rho(n_2, N_1)$, have large positive values (see Table 1); the correlation coefficient varies from -1 to 1 and is

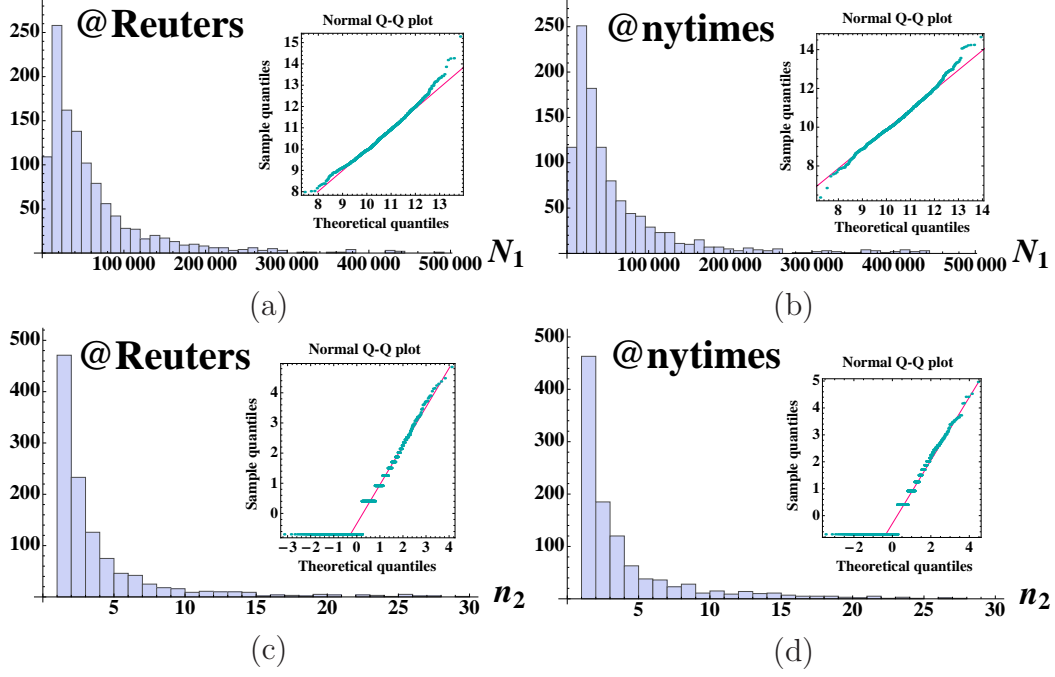


Figure 3: (Color online) The histograms of the retweet rates and the normal Q-Q plots of the logarithms of the number of followers and the retweet count in the second generation. (a) and (c) are of N_1 and n_2 for @Reuters. (b) and (d) are of N_1 and n_2 for @nytimes.

calculated by

$$\rho(n_2, N_1) = \frac{\langle n_2 N_1 \rangle - \langle n_2 \rangle \langle N_1 \rangle}{\sqrt{\langle n_2^2 \rangle - \langle n_2 \rangle^2} \sqrt{\langle N_1^2 \rangle - \langle N_1 \rangle^2}}. \quad (8)$$

Our result that the retweet rate β_2 obeys a lognormal distribution is plausible because lognormal distributions have the reproductive property, *i.e.* for two stochastic variables X_1 and X_2 which obey lognormal distributions,

$$p(\ln X_1) * p(\ln X_2) = \mathcal{N}(\mu_1, \sigma_1^2) * \mathcal{N}(\mu_2, \sigma_2^2) = \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2), \quad (9)$$

where $*$ stands for the convolution.

4. Model analysis: estimation of the diffusion range

From the model which we introduced above, we can estimate how much of the retweet rate $\beta^{\text{th}}(m)$ is required to reach the m th generation on average

	@Reuters	@nytimes
Number of seed tweets	1352	1140
Period	from Jun. 26, 2012 to Aug. 9, 2012	from Jun. 19, 2012 to Aug. 18, 2012
N_0	1 940 477	5 882 680
$\langle n_{\text{RT}} \rangle$	74.0	97.3
$(\langle \beta_1 \rangle, \langle \beta_2 \rangle)$	$(3.27 \times 10^{-5}, 7.90 \times 10^{-5})$	$(1.43 \times 10^{-5}, 8.52 \times 10^{-5})$
$(\langle N_1 \rangle, \langle n_2 \rangle)$	(75 276, 4.46)	(76 533, 4.54)
$\rho(n_2, N_1)$	0.468	0.481
$\langle c_1 / \sum_{f=1}^{n_1} k_f \rangle$	0.0471	0.0470
Fitting parameters for $\beta_1 = e^{\omega_1} + \delta_1$ $p(\omega_1) = \mathcal{N}(\mu_1, \sigma_1^2)$	$\mu_1 = -10.51$ $\sigma_1^2 = 0.6$ $\delta_1 = 0$	$\mu_1 = -11.51$ $\sigma_1^2 = 0.77$ $\delta_1 = 1.0 \times 10^{-6}$
Fitting parameters for $\beta_2 = e^{\omega_2} + \delta_2$ $p(\omega_2) = \mathcal{N}(\mu_2, \sigma_2^2)$	$\mu_2 = -9.65$ $\sigma_2^2 = 0.68$ $\delta_2 = -1.0 \times 10^{-5}$	$\mu_2 = -9.51$ $\sigma_2^2 = 0.69$ $\delta_2 = -1.0 \times 10^{-5}$
Fitting parameters for $N_1 = e^{\omega} + \delta$ $p(\omega) = \mathcal{N}(\mu, \sigma^2)$	$\mu = 10.64$ $\sigma^2 = 0.99$ $\delta = 0$	$\mu = 10.58$ $\sigma^2 = 1.04$ $\delta = 2000$
Fitting parameters for $n_2 = e^{\omega} + \delta$ $p(\omega) = \mathcal{N}(\mu, \sigma^2)$	$\mu = 0.48$ $\sigma^2 = 1.12$ $\delta = 0.5$	$\mu = 0.49$ $\sigma^2 = 1.23$ $\delta = 0.5$

Table 1: Data of tweet diffusions from @Reuters and @nytimes. The angular bracket $\langle \dots \rangle$ stands for the sample average.

and the average of the total number of retweets, $\langle n_{\text{RT}} \rangle$, for given parameters. In this section, we restrict ourselves to the case where the retweet rates β_g are independent of each other and their averages have a common value $\langle \beta \rangle$.

According to Eq. (5), the average of the number of retweets in the m th generation reads $\langle n_m \rangle = N_0 \bar{k}^{m-1} \langle \beta \rangle^m$. Then we have the threshold for the retweet rate where the diffusion reaches the m th generation on average, *i.e.* $\langle n_m \rangle \geq 1$:

$$\beta^{\text{th}}(m) = \left(N_0 \bar{k}^{m-1} \right)^{-\frac{1}{m}} = N_0^{-\frac{1}{m}} \bar{k}^{\frac{1}{m}-1}. \quad (10)$$

The behavior of Eq. (10) is exemplified in Fig. 4(a); in the case of the seed accounts which we investigated, the tweets diffuse up to the second or the third generation (see $\langle \beta_1 \rangle$ and $\langle \beta_2 \rangle$ in Table 1). While we employed the mean value of n_m in the definition of the threshold, it is also plausible to consider the median of n_m instead.

For a given range M of the diffusion, it is straightforward to calculate the average of the total number of retweets,

$$\langle n_{\text{RT}} \rangle = \sum_{g=1}^M \langle n_g \rangle = N_0 \langle \beta \rangle \sum_{g=0}^{M-1} (\langle \beta \rangle \bar{k})^g = N_0 \langle \beta \rangle \frac{1 - (\langle \beta \rangle \bar{k})^M}{1 - \langle \beta \rangle \bar{k}}. \quad (11)$$

The behavior of Eq. (11) is exemplified in Fig. 4(b); it shows that $\langle n_{\text{RT}} \rangle$ is not very sensitive to the diffusion range M in the case where $\langle \beta \rangle \bar{k}$ is small.

5. Conclusion and Discussion

We decomposed the diffusion of the daily tweets into the dynamics along the generations of the followers; the dynamics of retweet activities can be modeled as a random multiplicative process with respect to the generation. We directly observed the multiplicative factors from the actual data of Twitter and confirmed that our model is indeed plausible. We found that the multiplicative factors roughly obey lognormal distributions. The important point about our model is that the diffusion occurs owing to the repetition of cooperative activities along the followers and thus, as far as we know, it belongs to a different class of information diffusion compared to the ones in the literature [9, 10, 11]. We also believe that Twitter is not the only network where such a diffusion process occurs.

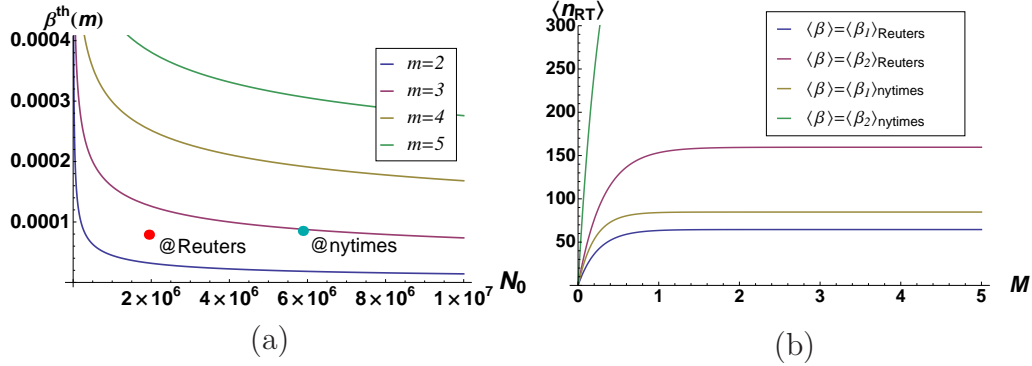


Figure 4: (Color online) (a) Threshold where the diffusion reaches the m th generation on average. We set $\bar{k} = 500$. The points are for @Reuters and @nytimes in the case where we assumed $\langle \beta \rangle = \langle \beta_2 \rangle$. (b) The average of the total number of retweets $\langle n_{\text{RT}} \rangle$ as a function of the diffusion range M for various values of the average retweet rate $\langle \beta \rangle$. We set $\bar{k} = 500$ and plotted the cases where $\langle \beta \rangle$ equals $\langle \beta_g \rangle$ of @Reuters and @nytimes. We set the values of them for N_0 , respectively.

The model of the present paper is a minimal model. In order to estimate the behavior of diffusion precisely, the approximation of the tree structure is obviously too rough; it is necessary to embed the information about the rate of over-counting and the heterogeneity of the network. We also neglected the correlation between the retweet rates. We will consider its effect in a future study which may be published elsewhere.

For the accuracy of the data, we sample the data of the news accounts only in the present paper. If Twitter API were updated and we tried a different way of sampling the data, we would be able to analyze the behaviors of many other accounts. Then we can proceed to a more quantitative analysis; for example, we would be able to measure the range of diffusion for each diffusion process.

6. Acknowledgments

The author thanks Naomichi Hatano, Tomotaka Kuwahara, and Tomohiko Konno for many useful discussions and Chris Wiggins for valuable comments. This work was partially supported by the Aihara Innovative Mathematical Modelling Project.

References

- [1] A. Java, X. Song, T. Finin, and B. Tseng: Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (ACM) (2007) 56-65.
- [2] H. Kwak, C. Lee, H. Park, and S. Moon: Proc. of 19th Int. World Wide Web Conf. (WWW) (2010) 591-600.
- [3] B. Krishnamurthy, P. Gill, and M. Arlitt: Proc. of the 1st workshop on Online social networks (ACM) (2008) 19-24.
- [4] J. Yang and S. Counts: Proc. of the Fourth International AAAI Conf. on Weblogs and Social Media (2010).
- [5] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi: Proc. of the Fourth Int. AAAI Conf. on Weblogs and Social Media (2010).
- [6] B. Suh, L. Hong, P. Pirolli, and E. H. Chi: Proc. of the IEEE Second International Conference on Social Computing (SocialCom) (2010) 177-184.
- [7] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer: WOSN'10 Proc. of the 3rd Conf. on Online social networks (2010) 3.
- [8] Note that the word retweet is sometimes used for two different meanings in the literature. The retweet button was first introduced at the end of 2009. Retweeting used to be simply a name of custom on Twitter to transfer the tweet of another user; it is called *informal retweet* nowadays, while the retweeting by clicking the retweet button is called *formal retweet*. Galuba *et al.* [7] also introduced a model of tweet diffusion in a very different manner from ours, but they limited themselves to the URL-embedded tweets and counted the informal retweets, whereas we analyze tweets in general and count the formal retweets in the present paper.
- [9] F. Wu and B. A. Huberman: Proc Natl Acad Sci USA **104**(45) (2007) 17599-17601.
- [10] D. M. Wilkinson: Proc. of the 2008 ACM Conf. on E-Commerce (2008) 302-309.

- [11] M. Yan and M. Gerstein, PLoS One **6** (2011) 5, e19917.
- [12] <https://dev.twitter.com/docs/api>, <https://dev.twitter.com/docs/streaming-api>
- [13] As long as the tweet owner is a public user, anyone can read the tweet and any user has the right to retweet, and therefore the chain of retweets among the followers is not the only way in which tweets diffuse. We do not treat such other processes in the present paper; we only counted the formal retweets [8] by non-private users because we believe that it gives the major contribution to the daily tweet diffusion.
- [14] We omitted the data with more than 800 retweets because Twitter API seems to fail to count the retweets correctly in such cases.
- [15] M. B. Wilk and R. Gnanadesikan, Biometrika **55** (1968) 1-17.